12. Visual environment and interlocutors in situated dialogue

Sarah Brown-Schmidt
University of Illinois at Urbana-Champaign

Abstract

Face-to-face conversation is often considered the most basic form of language use, as it was likely a dominant mode of communication as languages evolved, it is often the primary form of language input during children's language acquisition, and it is a dominant mode of adult communication today. Conversational language differs in important ways from the language traditionally studied in psycholinguistics; thus, characterizing language processing in conversation is essential if models of language understanding are to extend to this most basic form of language use. This chapter will examine key features of language comprehension in conversation, and will highlight the role of the visual environment in establishing joint domains of reference. Unlike in non-interactive settings, in conversation language is jointly created by conversational partners who hold different, but partially overlapping representations of the relevant context. Understanding if and how partners appreciate their partner's perspective has emerged as a central question in this domain.

Conversational language differs in important ways from the language traditionally studied in psycholinguistics. Conversation is situated in a context that is relevant to the language itself. This context may constitute the physical environment of the interlocutors, their shared history, the dialog itself, or some other combination of jointly established contextual knowledge. Conversation is also critically shaped by the fact that it involves the participation of at least two individuals. The result of multiple participation is that comprehension and production processes occur together in the moment; the ubiquity of split turns, in which one dialog partner finishes the other's utterance, is a prime example of this (Poesio & Rieser, 2010; Purver & Kempson, 2004). Thus, language in conversation is jointly created (Clark, 1992) and as such cannot be fully understood if processes of only one of the participants are isolated and studied. Other central features of conversation that are often absent in laboratory speech (unless they are the object of study) include the presence of disfluency (Arnold, Tanenhaus, Altmann, & Fagnano, 2004), gesture (Clark & Krych, 2004; Schegloff, 1984; Levy & McNeil, 1992), backchannels and other forms of feedback (Clark and Schaefer, 1989; Roque & Traum, 2008). Each of these features makes the *form* of language in conversation different than the form typically studied in standard psycholinguistic paradigms. While methodological innovations including the Visual World Paradigm (Tanenhaus, et al., 1995; also see Cooper, 1974; Pechmann, 1989, as well as Spivey & Huette, this volume, and Pyykkönnen & Crocker, this volume) support the study of contextualized language, implementations of this paradigm often lack the fully fledged interactivity that is typical of natural conversation.

Unlike language use in conversation, laboratory language is typically constrained in various ways in order to carefully control the language under study. Often, language production and language comprehension processes are studied separately, thus speakers produce and understand language in isolation. In language production studies, the speaker is typically not the originator of the ideas she speaks; instead many methodologies require speakers to repeat back a sentence, or to describe aspects of a scene selected by the experimenter. While these methodological controls afford consistency of productions across subjects, they

excise from the language production process most, if not all, of the first, and perhaps most important step in language production: the formulation of the to-be-communicated message (see Konopka & Brown-Schmidt, 2014). Similarly, in language comprehension studies, listeners are generally asked to interpret a series of unrelated sentences. Often these sentences are pre-recorded (or pre-typed, in the case of studies of reading), and thus unlike conversation, they are not created in-the-moment for that particular addressee.

Consider the following examples. The first is a series of linguistic stimuli presented to participants in an experiment by Trude and Brown-Schmidt (2012). In this study, participants listened to ~700 instructions like those in (a), one after the other. Compare these linguistic stimuli with the language in (b), which is an excerpt of a conversation from Brown-Schmidt and Tanenhaus (2008). In this study, pairs of naïve participants (1 and 2) worked together to arrange blocks in a visual display.

(a) *Click on tag.*
Click on back.
Click on wig.

(b)
1. *umm pushed down far down on to the top of the green is a little blue one*
2. *blue square?*
1. *yeah blue square*
2. *got it*
1. *ok*
2. *alright um…now…thuh um…go left from the blue square*
1. *yeah*
2. *there should be four spaces between that…and a penguin*
1. *a penguin*

In both experiments, a critical dependent measure was the eye movements that addressees made as they resolved lexical competition between cohort competitors in the visual display. For example, Trude and Brown-Schmidt examined fixations to

a picture of a bag when addressees interpreted the word *back*. In contexts in which both "bag" and "back" are potential referents, the shared initial phoneme results in competition between the two words, as evidenced by an initial rise in the likelihood of a fixation to both referents (Allopenna, et al., 1998). Similarly, Brown-Schmidt and Tanenhaus (2008) examined fixations to a picture of a pencil when addressees interpreted the word *penguin*. The form of the language in (b) is arguably more typical of every-day language use, yet the language in (a) is more typical of the scripted stimuli used in psycholinguistic research. The question, then, is whether these differences matter for the phenomena of interest.

In this chapter, I argue that the central phenomenon of interest in research on language processing is (or should be) how language is processed in everyday settings. Certainly, procedures such as reading, or listening to pre-recorded announcements are everyday behaviors. However, neither is more canonical, prevalent, or basic as everyday conversation. For example, the American Time Use Survey (US Dept of Labor, 2010) reports that in 2009, Americans spent approximately 42 minutes a day devoted to socializing and communicating—this was more than three times as much time spent on phone calls, mail and e-mail combined (12 minutes)[1]. While the quantity of all of these activities paled in comparison to television watching (169 minutes), I argue that TV is a less basic form of language use given that it is a modern development and not ubiquitous globally (at least not at such high quantities). A further consideration is that not all languages are written and even in modern societies, some proportion of the population is illiterate: The US national estimate for adults lacking "basic prose literacy skills" was 14 percent (2003, National Center for Education Statistics). Unlike television and text, spoken language is the form of speech that infants learn

[1]These data come from a 15 minute telephone survey of civilian adults over age 15. These values include only the primary activity and do not include any co-occurrent activity. Socializing and communicating is defined as "face-to-face social communication and hosting or attending social functions." Thus, uses of language at the same time as another activity (e.g., cooking) is not included in this estimate, and likely accounts for the intuitively low estimate. In 2013 the values were 43 minutes per day for socializing and communicating, 9 minutes for phone, mail and email combined, and 166 minutes for TV.

to speak their language from. Exposure to face-to-face language preserves the loss of non-native consonants, but exposure to pre-recorded audio or video does not (Kuhl, Tsao, & Liu, 2003). Similarly, exposure to infant-directed media does not increase vocabulary learning and is significantly less helpful than face-to-face interaction in the acquisition of new words (DeLoache, et al., 2010).

If we grant, then, that conversational language is the most basic form of language use worldwide and across the lifespan, we must consider whether the results of investigations of language in other forms, such as reading, listening to scripted sentences as in (a), etc. will extend to conversation. Answering this question will require the examination of language processing in conversational settings. The results of this research will indicate which findings from laboratory settings do and do not generalize to everyday conversation, as well as the boundary conditions that determine whether a finding will generalize. Studies of conversation also provide opportunities to make basic observations about mechanisms of language processing in every-day settings, which in turn, can be tested in more controlled laboratory settings, or in blended experiments that combine features of controlled experiments with features of natural conversation. This pairing of naturalistic studies with more tightly controlled traditional experiments will afford a more complete understanding of the mechanisms of everyday language processing than could be had from traditional laboratory studies alone.

This chapter focuses on interactive conversation, and explores how conversational partners, also known as interlocutors, coordinate meaning in conversation. In particular, I focus on the problem of establishing a *referential domain*, within which referring expressions are produced and interpreted. This chapter focuses on the way in which referential domains are shaped in conversation, and the implications this has for language understanding. In doing so, I lay out the case for the claim that insights gained from the study of conversational language are likely to be different in important ways than the insights that can be obtained by studying the scripted language typical of laboratory investigations. In the final section, I outline two alternative views of how referential domains might be constrained in conversation.

Referential Domains

All language is understood with respect to a context, whether it be the context of a conversation, the context of a paragraph in a book, or the context of a psycholinguistic experiment. The domain within which referring expressions are produced and interpreted is known as the referential domain. Classic research on reference in context demonstrates the sensitivity of referring expressions to the contents of the referential domain (Olson, 1970; Osgood, 1971). Imagine, for example, we wish to refer to Nabokov's novel, *Pale Fire*. In the context of a large library, to refer to the book, one would have to first mention both the title and the author in order to establish a referential domain within which the expression, *the book*, could be interpreted. In a context with only a few books, successful reference could be established by mentioning the color of the dust jacket, e.g., *the blue book*. In a face-to-face conversation, a pointing gesture could be used to further narrow the referential domain, allowing the speaker to use a pronoun, as in *Is <u>this</u> a good read*?

This dependence on context places a premium on understanding what the relevant context, or referential domain, is when understanding language. In the words of Lila and Henry Gleitman, "A picture is worth a thousand words, but *that's the problem*" (Gleitman & Gleitman, 1992, emphasis added). The world is always a source of context; what is unclear is which part of the world is the relevant part. How is it that we dice up the world into smaller referential domains? This is a problem that interlocutors appear to seamlessly and effortlessly solve, yet one that is a serious problem for theories of language use. This chapter explores two ways in which referential domains are established and circumscribed in conversation. The first is through the establishment of joint attention. The second is through representations of the perspective of one's dialog partner.

Joint Attention

The ability of communication partners to coordinate is often viewed as a prerequisite to successful communication (Clark, 1996; Clark & Brennan, 1991); when attention is coordinated, communication is thought to improve (Brennan, et

al., 2008; Richardson & Dale, 2005). According to one theory, coordinated attention during conversation improves communication by minimizing joint collaborative effort (Clark & Brennan, 1991; Gergle, Kraut, & Fussell, 2004a,b). If attention is coordinated, then speakers and listeners will produce and understand language with respect to the same context, and thus both production and interpretation processes should be more efficient.

Interlocutors can coordinate attention in a variety of ways, including gaze (Richardson & Dale, 2005; Richardson, Dale, & Kirkham, 2007), gesture (Bangerter, 2004; Clark & Krych, 2004), and actions in a joint workspace (Brennan, 2005). Furthermore, coordination of phonetic form (Pardo, 2006), syntactic form (Levelt & Kelter, 1982; Branigan, Pickering, & Cleland, 2000; Haywood, Pickering & Branigan, 2005; Reitter & Moore, 2007; Reitter, Moore, & Keller, 2006), and task schemas (Garrod & Anderson, 1987; Schober, 1993), as well as mimicry and coordination of body movements and posture (Chartrand & Bargh, 1999; Kendon, 1970) also emerge during dialog and may further reflect interlocutors' representational alignment (see Pickering & Garrod, 2004). Consistent with the view that interlocutors coordinate in order to minimize collaborative effort (Clark & Schaefer, 1989; Clark & Wilkes-Gibbs, 1986), partner mimicry effects may increase rapport and facilitate communication (LaFrance, 1979; LaFrance & Broadbent, 1976; Chartrand & Bargh, 1999; Richardson & Dale, 2005), even in human-computer interactions (Bailenson & Yee, 2005).

The present focus is on how coordination of attention can be used to support successful communication in conversation by establishing joint referential domains. In particular, this section focuses on gaze, gesture, and action as mechanisms for this coordination. See Chapter 9 of this volume (Knoeferle), for an in-depth treatment of the role of the visual context in sentence comprehension.

Gaze

Shifts in gaze are linked to shifts in attention, and the direction of fixation is typically taken as an indicator of the direction of attention (see Irwin, 2004 for discussion of this assumption). Gaze is also an important source of social and

attentional information in human development and learning. From infancy, humans are sensitive to the direction of adult gaze (Morales, Mundy, & Rojas, 1998; Morales, et al., 2000; Caron et al., 2002; Deák, Flom & Pick, 2000; Scaife & Bruner, 1975), and 18-month-old infants can use speaker gaze and gestures to learn the name for a novel object (Baldwin, 1991; 1993; also see Moses, Baldwin, Rosicky, & Tidball, 2001). Adults, too, can use the information about speaker gaze to learn novel words in an unfamiliar language (Yu, Ballard & Aslin, 2005).

The role of gaze extends beyond that of an attentional cue and a source of information during language acquisition. Gaze can also play an important role in on-line language processing, and it serves as a reliable indicator of communicative success.

Imagine a situation in which a dialog partner glances to the side and remarks, *That's neat!* In this context, the addressee can use the direction of the speaker's gaze to narrow the referential domain to a subset of entities in the general direction of the speaker's gaze, thus facilitating interpretation of what would otherwise be an underinformative expression. Hanna and Brennan (2007) demonstrated that addressees do just that. Participants in their experiment interpreted expressions like *the blue circle with five dots on it*, in contexts that contained two blue circles, one with five dots and one with six dots, and several objects of other colors. In a visual scene such as this one, the expression is temporarily ambiguous between the two blue circles. The ambiguity is resolved linguistically at the point-of-disambiguation (Eberhard, et al., 1995), which in the context of the task is the word *five*. Hanna and Brennan asked if speaker gaze could allow addressees to resolve this ambiguity earlier than the point-of-disambiguation. They hypothesized that addressees might use the direction of the speaker's gaze to narrow the referential domain to a subset of the task context. To test this hypothesis, they created situations in which pairs of naïve participants were seated on opposite sides of a visual display in which the objects were lined up in a row, between the participants.

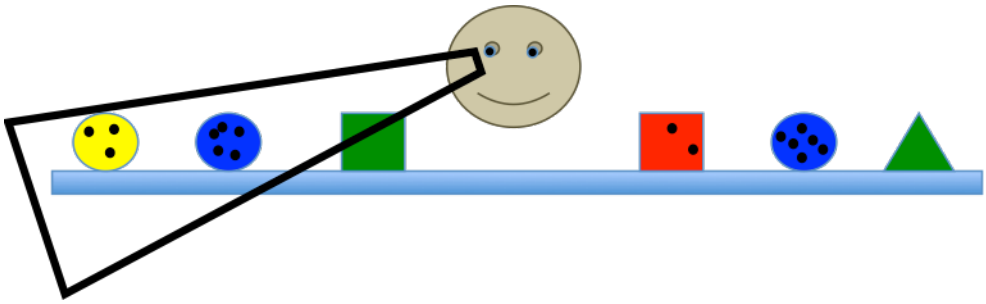


Figure 1. Gaze directs attention, narrowing the referential domain. Example display adapted from Hanna and Brennan (2007), Figure 1. Scene contains (left to right): yellow circle-3 dots, blue circle-5 dots (the target), green square, red square-2 dots, blue circle-6 dots (the competitor), green triangle. The speaker gazes to her right, excluding the competitor from the referential domain.

In one condition, two blue circles were on opposite sides of the display (see Fig. 1). Speaker gaze typically precedes reference to named objects by about 900 ms (Griffin & Bock, 2000), thus as speakers in this experiment prepared to say *the blue circle…*, their gaze was likely to be a reliable cue to speaker meaning. Addressees were highly sensitive to this cue, and within the first few hundred milliseconds after the onset of the adjective *blue*, fixations to the target referent rose quickly, with addressees identifying the gazed-at blue circle as the intended referent. This finding demonstrates that gaze is a source of information that addressees use to reduce referential ambiguity. Gaze narrowed the referential domain to objects in the direction of the speaker's fixation. Within this limited referential domain, the referring expression was no longer ambiguous.

This result is in line with other findings that giving one dialogue partner information about the other partner's gaze (real or simulated) can facilitate task performance. For example, Brennan, et al. (2008; also Neider, et al., 2010) asked pairs of eye-tracked participants to complete spatial tasks in which they had to search for a target in a scene with competitor (e.g., an O in the context of many Qs). Providing pairs with information about where their partner was looking (a live feed from the eye-tracker was displayed on their screen) speeded target identification—even more so than when partners could talk to each other, or even when they could talk and see their partner's gaze. Again, it seems that information about a partner's gaze was used as a tool to narrow the relevant domain, in this case, in a visual

search task. The fact that gaze could be such a powerful communicative tool—even more powerful than language itself—provides evidence that modes of communication other than spoken language play a key role in communicative processes.

Gaze is not only a *source* of information for communicative partners, but it also provides a good *measure* of the degree of coordination in conversation. For example, Richardson, et al. (2007; also see Richardson & Dale, 2005; Richardson, Dale, & Tomlinson, 2009) monitored the eye movements of participants as they conversed about the painting *Nature Morte Vivante* by Salvador Dali. Prior to their discussion, the speakers heard either the same or a different informational passage about Dali (either about the painting, or about Dali himself). When partners had the same background experience—the same common ground (Clark & Marshall, 1981)—their gaze during the subsequent conversation was significantly more coordinated. That is, when one partner looked at an element of the painting, the other partner looked too (with some lag of course, as speech-related gaze precedes speech and comprehension-related gaze follows it). Similarly, Richardson and Dale (2005) recorded speakers talking about a TV show as they gazed at images of key cast members. Later, a group of listeners listened to the recordings while viewing the images of the cast. Again, gaze proved a powerful indicator of communicative success: the more highly correlated speaker and listener gaze was, the more successful the communication (as evidenced by listeners' answers to comprehension questions). Thus in this task, when attention (measured by the direction of gaze) was similar, communication was more successful, likely in part due to similar referential domains.

These insights about human attention and referential domains are relevant not only to theories of language processing, but also to the field of artificial intelligence. The virtual human, Max, developed by the artificial intelligence group at the University of Bielefeld, Germany, is one good example. Max is an incredibly convincing virtual dialog partner. He makes use of information about a human's gaze and pointing gestures to assess their focus of attention. In doing so, Max is able to establish joint attention with the human communicative partner, and increase

fluidity of the interaction (see Pfeiffer-Leßmann, & Wachsmuth, 2009; Wachsmuth, 2008). Information about the human partner's attention, in combination with emotion simulation, intention recognition, and the ability to give feedback in conversation (Becker-Asano & Wachsmuth, 2010; Wachsmuth, 2008), make the experience of interacting with Max seem virtually real.

Actions and Gesture

In conversations about entities in the co-present world, referential domains can further be circumscribed by body movements, such as pointing gestures, and actions in the environment.

During a lengthy conversation, the partners' conversational history serves as a resource for information that can be used to circumscribe domains. Take, for example, the dialog presented in example (b). In that study, Brown-Schmidt and Tanenhaus (2008) examined the interpretation of expressions like *the penguin* in the context of both the target referent (a block with a picture of a penguin on it), and a competitor referent (a block with a picture of a pencil). They compared expressions that were produced during the course of a ~2 hour conversation in which partners worked together to arrange blocks in the same pattern on their respective game boards. In typical studies of speech perception using the visual world paradigm (Tanenhaus, et al., 1995), both a penguin and a pencil would be present on the display, and would thus both be potential referents. Those studies typically find that shortly after the onset of the word *penguin*, the addressee launches fixations to both the penguin and the pencil, with roughly equal likelihood, until disambiguating phonetic information is heard (Allopenna, et al., 1998). In these studies, various sources of information, such as subphonemic coarticulatory information (Dahan, et al., 2001), information about a particular speaker's referring tendencies (Creel, Aslin, & Tanenhaus, 2008), and information about a particular speaker's vowel shift (Trude & Brown-Schmidt, 2012) all modulate this process.

How is this type of lexical competition resolved during conversation? To address this question, Brown-Schmidt and Tanenhaus first examined interpretation of these expressions for language outside the context of the conversation itself. To

do this, they had the experimenter refer to various game pieces on the board as in *Look at the penguin, ok… Look at the lamp….* In this context, the typical cohort competition effect was replicated, with an early rise in fixations to both alternatives. In contrast, reference to the exact same game pieces made during the course of the conversation elicited no detectable competition effects. Addressees were no more likely to look at competitors than unrelated blocks (e.g., a candle when interpreting *candy*). Instead, in most cases listeners had already focused visual attention on the target prior to the referring expression, and did not direct attention away from the target when hearing a word that was temporarily consistent with a competitor. Further, in situations where listeners were not already fixating the target prior to the target word, fixations to the target rose rapidly following target word onset, and there was no detectable competition effect.

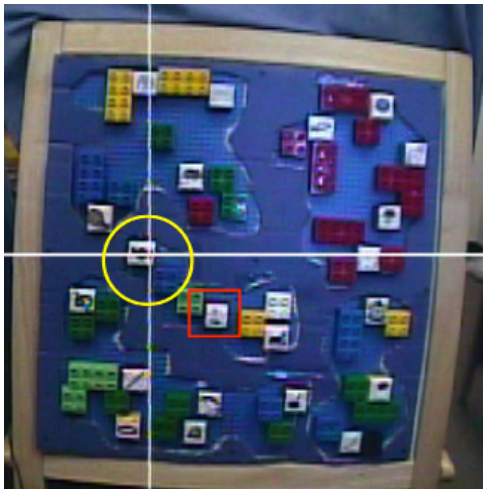


Figure 2. Task constraints narrow the referential domain: Screenshot from Brown-Schmidt and Tanenhaus (2008). Participant is fixating the "candy", indicated by white crosshair. The yellow circle indicates the possible referential domain; the competitor, "candle" (highlighted by a red square) is outside the hypothesized referential domain.

This effect was interpreted as a referential domain effect. Brown-Schmidt and Tanenhaus argued that the interlocutors constrained their referential domains to such small areas of the board that the expressions were no longer ambiguous: that is, the candle (when interpreting *candy*) was simply not a competitor. Further

analyses examined how the domains came to be constrained. While up to 57 potential referents were on the board at any given time, speakers and addressees only considered those that had been mentioned recently, that were relevant to the task, and were in close physical proximity to the last mentioned object. Similar task-based constraints have been found to constrain referring in other task-related conversations (Beun & Cremers, 1998; also see Landragin, 2006), suggesting these effects are not limited to the particular task used in this study.

Lexical competition during spoken word recognition can be attenuated by other constraints as well, including semantic information (Barr, 2008), talker preferences (e.g., if one talker always says *candy*, and a different talker always says *candle*, Creel, et al., 2008), and structural priming of verbs (Thothathiri & Snedeker, 2008). Possible actions in face-to-face conversation can also constrain domains. For example, Hanna & Tanenhaus (2004) demonstrated that during a task-based conversation in which a confederate (someone pretending to be a genuine participant) was following a recipe to bake a cake along with a participant, that the confederate's ability to reach to certain items in the workspace constrained which items were considered relevant. On critical trials, the context contained two boxes of cake mix, one of which the confederate could reach with her hands, and one of which she could not, and she asked the participant to *put the cake mix*…. In cases where the confederate's hands were empty, the expression was interpreted as referring to the cake mix that the confederate could not reach. In this case, the competitor was considered outside the referential domain because if the confederate had wanted that cake mix, she would have reached for it herself. In contrast, when the confederate's hands were full, both boxes of cake mix were considered. This result indicates that the referential domain is changed by the possible actions that could be performed in a situation.

Executed actions play other roles in conversation as well, including acting as a stand-in for language, and providing tangible evidence of understanding. Providing shared visual information as conversational participants complete a joint task affords the use of actions in the place of words. Clark and Krych (2004) found that listeners used pointing gestures and actions such as holding a block in a certain

location to demonstrate the listener's understanding during task-based conversation. When the joint workspace was hidden from the speaker, they observed that pairs tended to spend more time checking whether a previous action was correct or not. Pairs with visible workspaces also tended to use more deictic expressions, particularly expressions like *like this,* or *like that*—these expressions were frequently combined with gestures in which an action was demonstrated (e.g., does it go *"like that*"?). Similarly, Gergle, Kraut, and Fussell (2004b) asked one participant to instruct another participant on how to assemble a 4-piece puzzle on a computer, and manipulated whether the director saw a live view of the matcher's workspace. Having a view of the matcher's workspace changed how they partners completed the task. When the director could see the workspace, actions in the workspace took the place of talk. These actions established whether the matcher correctly understood or not, and as a result there were fewer verbal acknowledgments of having moved a piece when workspaces were shared.

An open question is how referential domains might be constrained in other situations. Take, for example, a discussion about a movie. During the movie itself, scenes change rapidly, and viewers may not keep track of even noteworthy changes to the objects in those scenes (Simons & Chabris, 1999). Scene changes result in rapidly changing object locations and viewpoints, and thus the relative location of potential referents. As a result, physical proximity, which is a constraint that features strongly in task-based conversation (Brown-Schmidt & Tanenhaus, 2008; Beun & Cremers, 1998; Hanna & Tanenhaus, 2004), may play less of a role. Further, segmentation of events may separate entities into separate referential domains, in both visual event perception (see Zacks, 2004) but also in the comprehension of narrative and possibly non task-based dialog (see Speer & Zacks, 2005; Greene, et al., 1994). These event representations may include expectations for unmentioned or unobserved changes (e.g., Altmann & Kamide, 2009). The semantic structure of complex events may also constrain domains. Physical and semantic constraints on the action of putting something "inside" narrows the domain of interpretation of a sentence like *Put the cube <u>inside the can</u>* to container-like goal locations that are physically compatible with the object to be put (Chambers, et al., 2002; also see

Dahan & Tanenhaus, 2004). Similarly, information about the indexical characteristics of event participants constrains the possible events they may engage in (Kamide, Altmann, & Haywood, 2003; Tesink, et al., 2008; van Berkum, et al., 2008). For example, in an analysis of event-related potentials to auditorily presented sentences, van Berkum, et al. (2008) found that listeners incorporated information about the age and gender of a talker into their interpretation of sentences. They found that mismatches between the talker and the information communicated by the sentence, such as *Every evening I drink some <u>wine</u> before I go to sleep*, spoken by a child elicited significantly larger N400 responses to the critical word *wine*, in comparison to a case where the speaker's identity was consistent with the information being conveyed (e.g., an adult).

These expectations based on semantic and indexical information are consistent with a view that interlocutors maintain detailed representations of contextual information. These partner-specific representations also include the *perspective* of one's partner, a topic we turn to next.

Perspective-Taking

In dialog, appreciating the knowledge state of one's interlocutor may be important for how the addressee understands language. Consider, for example, the excerpt (c) of dialog from the television show "Friends"[2].

(c)

Phoebe: *They don't know that we know they know we know! Joey, you can't say anything!*

Joey: *I couldn't even if I wanted too.*

In this exchange, the characters are discussing the mutual awareness of the fact that the characters Monica and Chandler are secretly dating. In the context of this TV series and this particular episode, the secretive dating and knowledge of this fact are well-established. As a result, the convoluted sentence, "*They don't know that we*

[2] From Season 5 Episode 14, "The One Where Everybody Finds Out". Transcript available from friends.wikia.com. See Cohen (2010) for discussion.

know they know we know!" becomes interpretable. Establishing a meaning for this sentence outside of a rich context is difficult because it involves the calculation of at least four embedded mental states (knowing of knowing of knowing of knowing). However, in the context of the show, the experience, knowledge and goals (i.e., to deceive) are salient, and interpretation of such multiply embedded statements comes fairly naturally.[3] According to one proposal (Brown-Schmidt, 2009a), these mental-state calculations should be facilitated in situations where the listener is participating in a live conversation, rather than passively listening (e.g., to the television), as a live interaction provides better opportunity to firmly establish what is and is not jointly known. The fact that sentences such as *They don't know that we know they know we know!* can be successfully interpreted when watching television may benefit from the build-up of information throughout the episode, as well as the viewer's familiarity with the show. Whether understanding of such sentences would be even easier in a live conversation, remains to be tested.

How is it that interlocutors compute mental states in such a way that they can be rapidly deployed for the purposes of understanding language—even language as convoluted as the above example? According to Clark and Marshall (1981) interlocutors establish enough mutual knowledge for the current purposes based on co-presence heuristics and assumptions about simultaneity of attention (among others). In Clark and Marshall's view, to establish a physically co-present object as part of the interlocutors' joint knowledge or *common ground*, interlocutors represent the fact that the given entity is mutually known if the entity and both interlocutors are co-present, and the interlocutors have evidence of each other's mutual attention to this entity. Other forms of co-presence include linguistic and cultural co-presence. On their view, information about the co-presence of entities and individuals is stored in rich, diary-like representations. This evidence for common ground varies in strength, such that some evidence offers a strong case to

[3] Rich semantic and contextual information can similarly assuage challenging syntactic constructions. The problematic syntactic structure in *The horse raced past the barn fell* is much easier to interpret when the lexical affordances are consistent with the syntactic structure, as in *Whiskey fermented in oak barrels can have a woody taste* (see McRae, Hare, & Tanenhaus, 2005).

assume common ground (e.g., we are both jointly looking at an object), whereas other evidence only provides weak support for common ground. In particular, Clark and Marshall (1978) suggest that linguistically mentioning something provides weaker evidence for common ground compared to immediate physical co-presence, because something mentioned in the past can only be considered common ground if the listener remembers what was said (among other criteria). These representations of joint knowledge, whether weak or strong, are accessed in the service of language. In example (c), the interlocutors would maintain information about each other's knowledge states, based on mutually observed events, such as visual evidence of an affair, or previous discussion of said affair. Access to these rich representations would then allow the character Phoebe in (c) to comment on another person's lack of knowing.

This view of common ground posits a central role for explicit memory processes in the use of mutual knowledge. Another view, proposed by Horton (Horton & Gerrig, 2005a, b; Horton 2007) posits that in addition to explicit recollection of joint experience, *I remember when Phoebe and I…*, common ground is formed on the basis of low-level associations between individuals and information. These associations could support use of language in a way that is sensitive to the common ground between individuals, without requiring that the sources of that information (jointly experienced events) be explicitly accessed from declarative memory during language use itself.

Understanding whether the representations underlying common ground are strictly episodic, diary-like representations, or whether there is an association-based component as well is a critical question for future research. The answer has implications for understanding if and when common ground could guide language processing. Common ground has the potential to play a powerful role in comprehension during conversation because it could constrain the domain of interpretation to information relevant to the dialog, based on the partner's perspective. For example, when interpreting an imperative, *Pick up your toy!*, the referent of *toy* is likely to be some entity mutually known to speaker and listener. If it was unknown to the speaker, she wouldn't refer to it, and if it was unknown to the

addressee, the speaker would need to provide more information if she wanted her command to be understood. By contrast, when interpreting a question like *What did you buy?*, the question is likely to be asking about something known to the addressee but not the speaker (see Brown-Schmidt, 2005). How does common ground constrain the domain of interpretation for language? Here we consider how establishment of common ground can influence referential domains, particularly focusing on common ground for physically co-present objects, and linguistically co-present entities. For a different view on the role of common ground in language see Chapter 11 of this volume (Barr).

Physical co-presence

Consider a situation in which two people sit face to face, across a table from each other. In face-to-face situations, the dialog partners have different physical viewpoints on a scene, resulting in different perspectives (Figure 3).

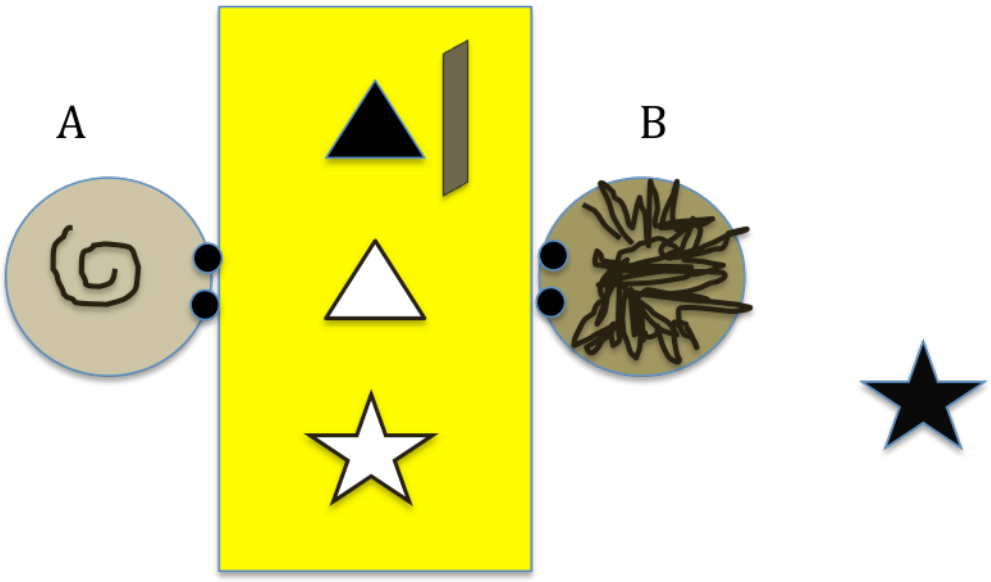


Figure 3. In face-to-face conversation, some entities are mutually visible (the white star and white triangle), and thus in common ground. Other entities might be occluded by a barrier (the black triangle), or located out of Partner B's sight (the black star), and thus in Partner A's privileged ground.

According to some views of language use, common ground is the basic context with respect to which language is produced and interpreted (Clark, 1992; 1996). On this

view, appreciation of which entities are and are not physically co-present would be a basic, and routine component of both language production and language comprehension processes. This would mean that while Partner A in Figure 3 sees two stars, he would not need to modify his expression to refer to the one on the table, as *the star* would be perfectly interpretable from Partner B's point of view. Similarly, if Partner B were to refer to *the triangle*, Partner A should understand her to mean the white triangle, as the black triangle is not visible from Partner B's perspective.

These predictions are not entirely consistent with the observed patterns of behavior in language production and comprehension. Instead, the literature suggests that common ground only partially constrains referential domains. Consider the case of Partner A's interpretation of B's expression, *the triangle,* in a sentence like *Pick up the triangle and move it next to the star.* Evidence from the analysis of eye movements in situations similar to this one show that addressees do sometimes consider the privileged (black) triangle (Keysar, Lin, & Barr, 2003; Hanna, Tanenhaus, & Trueswell, 2003), suggesting that information about what information is common and privileged is not an absolute constraint on the referential domain. However, it is a partial constraint: addressees in Partner A's perspective are significantly less likely to gaze at a privileged competitor compared to one in common ground (Hanna, et al., 2003; also see Heller, et al., 2008; Chambers & San Juan, 2008). In language production, speakers show sensitivity to the perspective of the addressee (Nadig & Sedivy, 2002) or addressees (Yoon & Brown-Schmidt, 2014), designing expressions that are consistent with the addressee's perspective at least part of the time. Taken together, these findings are consistent with constraint-based views of common ground (Brown-Schmidt & Hanna, 2011; Hanna, et al., 2003), which propose that common ground is one of many partial constraints on language processing.

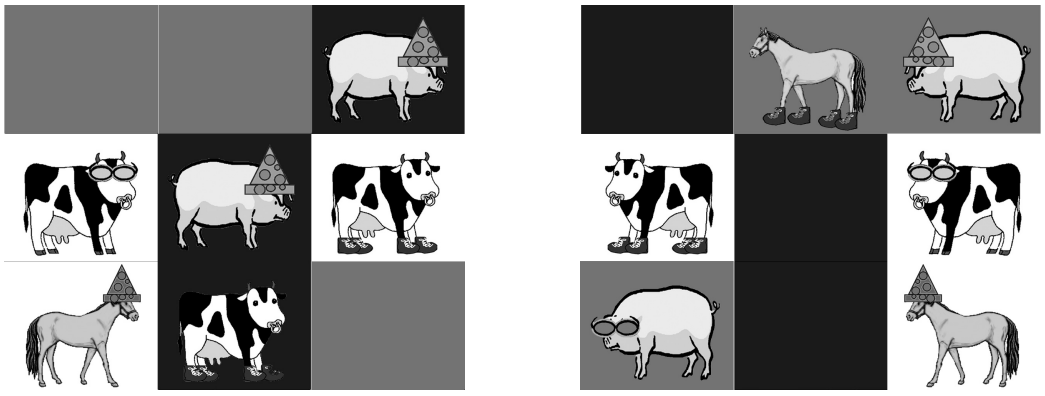
In the situation depicted in Figure 3, information about what is common or privileged is provided by visual cues in the context, what Clark and Marshall (1978) termed *physical co-presence*. This is the most typical type of situation studied in experiments on common ground. However, Clark and Marshall outlined another

scenario for the visual establishment of common ground, *delayed physical co-presence*. Imagine a situation in which Partners A and B jointly gaze at the white star, but then the star falls off the table, out of view. The fact that the white star had been established in common ground at one point would then allow reference to it after some delay. Little research directly investigates this source of information about common ground. In one study, listeners did not use delayed physical co-presence to guide referential processing (Ryskin, et al., 2014), though another study that used simpler displays and shorter delays did find sensitivity to previously-established physical co-presence (Ferguson & Breheny, 2012). Understanding the constraints on the use of delayed physical co-presence remains an important question for future work; considerations of the memory demands involved (Horton & Gerrig, 2005a; Rubin, et al., 2011) are likely to be relevant.

A different way of establishing common ground is through the use of language, that is, by mentioning new information to your dialog partner that was previously privileged. In the next section, we discuss evidence for how *linguistic* cues to common ground guide language processing.

Linguistic co-presence

For any two individuals, their beliefs and knowledge are necessarily non-identical. Thus, much of conversation involves exchanging information that was previously not mutually known. In this way, dialog partners take information that was previously privileged and make it *linguistically co-present*. How does linguistic co-presence compare to physical co-presence as a source of information about common ground? Clark and Marshall (1978) suggested that linguistic co-presence provides weaker evidence for common ground, in part because interlocutors have to remember what was mentioned, whereas physical co-presence (the immediate kind) is available in the here and now. The limited empirical evidence on this question, however, suggests that linguistic and physical co-presence may be comparable.



(a) (b)

Figure 4. Example scene from Brown-Schmidt, et al. (2008), from the experimenter's (a) and participant's (b) perspective. Displays are mirror-reversed. Animals in white squares are visually co-present, and animals in black/gray squares are visually privileged.

Brown-Schmidt, Gunlogson, and Tanenhaus (2008; also see replication by Ryskin, et al., 2014) examined the use of common ground as participants interpreted informational questions like *What's below the cow with shoes?,* given scenes like the one in Figure 4. In this example, the underlined portion of the question is temporarily ambiguous between asking about the animal below the cow with shoes (the target) and the animal below the cow with glasses (the competitor). However, the animal below the cow with glasses is already common ground (the horse with the hat). Thus, if participants can use physical co-presence to constrain the referential domain to things appropriate to ask questions about, then the question is disambiguated at the word *cow*. Brown-Schmidt, et al. (2008) compared this condition to a case in which the competitor was visually privileged, but linguistically mentioned prior to the critical question. The results were the same across the two conditions: When common ground (physical or linguistic) ruled out the competitor, participants began to look at the target more than the competitor shortly after the onset of the critical noun, *cow*.

These results suggest that, at least in some circumstances, physical and linguistic sources can provide equally good information about common ground. By contrast, failures to use delayed physical co-presence (Ryskin, et al., 2014) may be due to problems in maintaining and/or retrieving this information over a delay

period. Similarly, when linguistic information had been introduced a long time ago, memory failures may impair use of common ground (see Rubin, et al. 2011).

Joint Attention

A final consideration is that establishing common ground based on physical and linguistic co-presence requires assumptions about joint attention. If Partner A in Figure 3 had his eyes closed it would not be appropriate to assume that the white star and triangle were common ground. Similarly, if Partner A were to say *There's a black star behind you*, A could only consider the black triangle to be common ground if B showed some evidence of understanding the utterance. If B was listening to her iPod at a loud volume, or was distracted, etc., assumptions about simultaneity of attention to A's speech could not be made. These examples illustrate the importance of *grounding* joint knowledge. According to classic theories of dialog, information is only entered into common ground if both partners accept it. One way of doing this is by providing feedback, as in *Ok, there's a star behind me, thanks!*, which can provide varying amounts of information for whether something is common ground (see Clark & Schaefer, 1989; Roque & Traum, 2008; 2009).

Brown-Schmidt (2009b) found some evidence that partners are sensitive to the grounding process. In that study, participants brought visually privileged animals into common ground by mentioning them. Critically, the feedback that the experimenter gave was manipulated. On some trials, the experimenter used positive feedback, as in *Okay*. In other cases, the experimenter gave negative feedback, as in *Sorry, I didn't get that*. Then, participants interpreted a wh-question that was temporarily ambiguous between asking about the information that had been mentioned, and something that had not been discussed. Participants were significantly less likely to consider the mentioned competitor when the experimenter provided positive feedback after the competitor's identity was revealed, compared to a case where the experimenter provided negative feedback. This result suggests that feedback does in fact play a role in establishing common ground. However, whether fine gradients between different forms of feedback are used (Clark & Schaefer, 1989), is an open question. In some of the only work to

address this latter issue, Brown-Schmidt (2012) reported minimal differences between the following forms of feedback: *OK* (see d2a), repeats (d2b), and continuations of the discourse (d2c).

(d)

d1. Participant: I have a horse with a hat in my secret square.

d2a. Experimenter: *OK*

d2b. Experimenter: *Horse with hat.*

d2c. Experimenter: *So now pick up the triangle and….*

Whether larger effects might be observed in other circumstances remains to be explored.

Towards a model of domain circumscription

The previous sections outlined ways in which referential domains are circumscribed in conversational settings: Eye fixations can limit the referential domain to entities in the direction of the speaker's gaze (Hanna & Brennan, 2007), task demands can limit the domain to task-relevant or recently mentioned items (Brown-Schmidt & Tanenhaus, 2008; Beun & Cremers, 1998), and common ground can limit a domain to information either in or out of common ground, depending on utterance form (i.e., an imperative vs. an interrogative; Hanna, et al., 2003; Brown-Schmidt, et al., 2008).

These examples demonstrate that domain circumscription reduces competition from potential referents during interpretation of a referring expression. How exactly is this ambiguity eliminated? In this final section, I discuss two possible mechanisms for how domains might be circumscribed in conversation, based on the factors discussed above. The first possibility is that addressees maintain a single, attentionally-constrained referential domain. This account is contrasted with a view in which addresses maintain multiple independent (and potentially inconsistent) domains (see Heller, Parisien, & Stevenson, 2012 for a related view).

The first possibility is that linguistic, pragmatic and other information define a *single, attention-focused referential domain*. Consider the left panel of Figure 5.

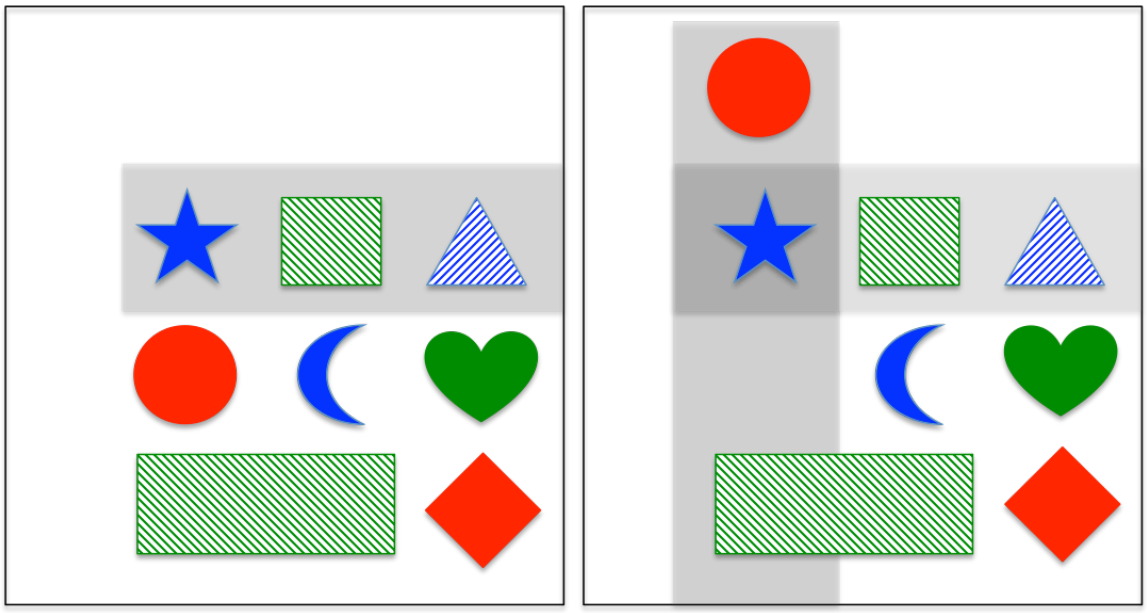


Figure 5. Example display. Left panel: Shaded area at top of display indicates hypothesized referential domain circumscribed by the word *above*. Right panel: Updated scene following movement of the circle. The shaded area at the left side of display indicates referential domain circumscribed by the word *left*. The star is in both referential domains.

Given this scene, if a subject were to hear the instruction, *Put the circle above the striped green square,* the referential domain during interpretation of the second referring expression (underlined) would be narrowed by the restrictions of the preposition *above* (see Chambers, et al., 2002), and a task-based constraint to not allow objects to overlap. The items in the domain would therefore be those entities with space above them—the star, the striped green square, and the triangle (Figure 5, left panel). On this view, during interpretation of the expression *the striped…*, only those three entities would be considered potential referents.

If language comprehension takes place with respect to a single referential domain, this begs the question of how domains are updated over time. If the mechanism of domain circumscription is attentional, listeners may have difficulty switching attention from one domain to the next, even after the first domain is no longer relevant (see Ryskin, et al., 2014 for a discussion of the costs of switching

between domains). In the above example, the referential domain was initially established as the items in the top row—the star, square and triangle (Figure 5, left panel). If the speaker subsequently gave an instruction to *Now put the moon to the left of…*, the word *left* would change the referential domain to be those items on the left side of the display—the circle, the star, and the rectangle (Figure 5, right panel). If switching attention from items in one domain to items in a different domain poses difficulties, then it should be difficult to interpret an expression that references an object not in the original domain (e.g., the rectangle), compared to an object which was included in both domains (e.g., the star). For example, the noun phrase *the star* in (e1) should be easier to interpret than *the rectangle* in (e2), because the star was in the previous referential domain.

(e1) *Now put the moon to the left of the star.*

(e2) *Now put the moon to the left of the rectangle.*

In the view discussed thus far, interlocutors maintain one referential domain at a time, and switch between domains as the conversation unfolds. How else might domain circumscription operate? An alternative possibility is that interlocutors maintain *multiple domains* in working memory, or task focus (Grosz & Sidner, 1986). On this view, different domains might include different entities, not all of which match the selectional restrictions of the incoming acoustic information at the time.

The advantage of a multiple-domains view is that it can account for why some sentences have multiple, conflicting domain restrictions. Consider the wh-questions examined in Brown-Schmidt, et al. (2008), such as, *What's below the cow with shoes*? In a sentence like this, the question is inquiring about something that must be in the addressee's privileged ground. However, interpreting the question requires understanding a definite reference to something in common ground (cow with shoes). Thus, within the same sentence, the referential domain must shift from privileged ground, to common ground, and then back again to privileged ground, in order for the addressee to answer the question. The speed with which such utterances are interpreted (Brown-Schmidt, et al., 2008; Brown-Schmidt, 2009b; Ryskin, et al., 2014) suggests that both common and privileged information are

available at once. Thus, effects like this one suggest that multiple conflicting domains might be active at one time.

Heller, et al. (2012) proposed a multiple-domains view of perspective-taking in which interlocutors maintain separate representations of common ground and privileged ground. They present data from studies of language production and comprehension in cases where speaker and listener have different perspectives. Heller, et al.'s findings suggest that these domains (common ground and privileged ground) are probabilistically weighted and combined together to guide language production and comprehension. Understanding how this probabilistic-combination view speaks to the problem of changing domains over time remains an important question for future work.

General Discussion

This chapter makes the strong claim that the object of study in language processing is, or should be, the most basic form of language use, which I claim is interactive conversation. Further, I argue that engaging in conversation changes the way in which language is processed in ways that are relevant to the phenomena under investigation. The bulk of this chapter explores these ideas by examining how referential domains are circumscribed during interactive conversation, and how interpretation of referring expressions is shaped by domain circumscription. In doing so, I outline two ways in which conversation narrows referential domains and speeds processing--through joint attention, and through perspective-taking. The goal of this final section is to summarize these findings, and discuss how they support the claim that conversation alters language processing in ways that are potentially relevant to the theoretical questions of interest.

The first source of domain circumscription I described is joint attention in conversation. In face-to-face conversation, interlocutors have access to a highly reliable cue to their partner's object of attention, gaze. Coordination of gaze in conversation can be used as a measure of joint attention, and as such, a reliable indicator of communicative success (Richardson, et al., 2007; Richardson & Dale, 2005; Richardson, et al., 2009). Gaze can also serve as an early cue to speaker

meaning during interpretation of a temporarily ambiguous referring expression (Hanna & Brennan, 2007), and can even allow the young child to infer the meaning of a novel word (Baldwin, 1991; 1993). Similarly, actions and gestures in a joint workspace not only focus attention, and improve communicative success (Clark & Krych, 2004), but more importantly, they can take the place of linguistic exchanges (Gergle, et al., 2004a,b), and alter the linguistic forms that speakers do use (Clark & Krych, 2004). While these physical cues are readily and naturally produced in conversation, they may be absent in some non-interactive forms of language use, such as reading, speaking in isolation, or listening to pre-recorded stimuli. The fact that these cues to joint attention are beneficial to processing suggests that comprehension processes may be impaired in non-interactive settings, a claim consistent with findings that communication suffers when these physical cues are eliminated (Clark & Krych, 2004; Gergle, et al., 2004b; Brennan, 2005). Further, the tendency for interlocutors to rely on cues such as gaze and actions (e.g., Hanna & Brennan, 2007; Clark & Krych, 2004) suggests that the constraints relevant to language processing (i.e., Trueswell & Tanenhaus, 1994) are qualitatively different in interactive settings. This therefore suggests that conclusions regarding which sources of constraint are central to language processing, and which are peripheral, must be qualified based on the mode of language use.

A second way in which domains are constrained in conversation is through representations of the perspective of one's dialog partner. In conversation, interlocutors form representations of common ground (Clark & Marshall, 1978; 1981) that are subsequently used to guide language processing in a manner dependent on the form of a given utterance. Whereas interpretation of a noun phrase in an imperative such as *Hand me the cheese* narrows the referential domain to entities in common ground (cheeses that we both know about), interpretation of a question such as *Where's the cheese?* narrows the domain to information in privileged ground—that is, the location of the cheese in question (Brown-Schmidt, et al., 2008; Brown-Schmidt, 2009b; Nurmsoo & Bloom, 2008). Representations of common ground are established through interactive processes of introducing and establishing information as shared (Clark & Schaefer, 1989; Roque & Traum, 2008;

2009; Brown-Schmidt, 2009b), with some sources of information for common ground providing stronger evidence of joint knowledge than others (Clark & Marshall, 1978). According to Brown-Schmidt (2012), representations of common ground vary in a gradient fashion depending on the amount of evidence for the assumption of mutuality. Consistent with the claim that common ground is gradient are findings that addressees are less likely to rely on representations of common ground in non-interactive settings where common ground is less well established (Brown-Schmidt, 2009a; Brown-Schmidt & Fraundorf, submitted). Referential understanding is generally impaired for non-interactive compared to conversational language (Branigan, Catchpole, & Pickering, 2011; Foxtree, 1999; Schober & Clark, 1989; Wilkes-Gibbs & Clark, 1992), and speakers show sensitivity to characteristics and naturalness of the dialog partner (Lockridge & Brennan, 2002; Kuhlen & Brennan, 2010; see Kuhlen & Brennan, 2013). Taken together, these findings provide strong initial evidence for the claim that language processing is different in interactive settings in ways that are relevant to theoretical conclusions of interest.

At the beginning of this chapter, I described two different experiments that used lexical competition (cohort) effects to examine language processing. The study by Trude and Brown-Schmidt (2012) used a non-interactive paradigm in which participants listened to approximately 700 trials over the course of 2 hours. On each trial, participants saw 4 pictures on the screen, and heard one of two pre-recorded voices, a male and a female, refer to one of the pictures, as in *Click on back*. The goal of the experiment was to examine if listeners could learn a particular characteristic of the male talker's voice, that the /æ/vowel in *bag* was raised to /eɪ/ only before /g/ (e.g., *bag* is pronounced /beɪg/). The results of this experiment showed that learners were, in fact, able to learn this second-order phonemic constraint and that as a result, when the male (but not the female) talker was speaking, fixations to the cohort competitor, the bag, were reduced (but not eliminated). This talker-specific effect was subtle, yet the result still obtained in this non-interactive paradigm. How might have the results have changed in interactive conversation?

The results reported by Brown-Schmidt and Tanenhaus (2008) suggest that during an interactive conversation, depending on the referential domain, the

competition that was eliminated by a learned vowel shift might not have been there to begin with. Outside the context of conversation, utterances produced by the experimenter elicited standard cohort competition effects (Allopenna, et al., 1998). How did conversation shape this effect? It eliminated it. Unlike the reduction in cohort competition seen in Trude and Brown-Schmidt (2012; and other non-interactive paradigms examining constraints on language interpretation; e.g., Dahan & Tanenhaus, 2004; Dahan, et al., 2001; Creel, et al., 2008; McMurray, et al., 2008), cohort competition was completely eliminated during interactive conversation. This effect was interpreted as a referential domain effect: pragmatic constraints narrowed referential domains to small, task-relevant areas of the workspace, with the result that lexical competition processes were largely eliminated.

What are the implications of these findings for our understanding of lexical competition resolution and language processing in general? The research reviewed in this chapter suggests that interactive conversational settings provide a rich source of information typically not available in the non-interactive, scripted settings routinely employed in psycholinguistics research. Conversations take place within a context that includes gaze, common ground, and a discourse history, and that dramatically constrains referential domains, improving the efficiency and success of language understanding. One implication is that problems typically seen as fundamental to language processing, such as the resolution of lexical competition, may be relatively minor problems in conversational settings where domains are routinely constrained (Brown-Schmidt & Tanenhaus, 2008), where talker identity and preferences limit the candidate referents (Creel, et al., 2008; Creel, 2014), or where physical cues such as eye gaze give away the speaker's referential intentions (Hanna & Brennan, 2007). Thus, a key contribution of research on interactive conversation is to suggest changes in the relevant focus of experimental work in language processing. In this case, the suggestion would be a shift away from a focus on how word recognition processes resolve competition between large numbers of candidate words, and towards a focus on understanding how interlocutors avoid lexical competition in the first place, e.g., through domain circumscription. In doing

so, it will become important to understand the mechanisms by which domains are constrained.

What does this mean for standard research paradigms? The results of research in conversational settings do show that language processing is altered by the context of conversation, and suggest that some problems that might seem significant in unnatural, decontextualized settings are more modest in interactive conversation. However, standard research paradigms—including the use of decontextualized language—play numerous essential roles in psycholinguistics research: These studies afford the incredibly well-controlled study of very specific aspects of language processing. Research on the way in which sentences are interpreted given verbs and noun phrases with particular affordances (e.g., Wilson & Garnsey, 2009; Garnsey, et al., 1997) would likely be near-impossible to do well in a completely unscripted conversational setting. Similarly, understanding how listeners learn features of a talker's native accent and use that information to guide on-line interpretation (e.g., Dahan, et al., 2008; Trude & Brown-Schmidt, 2012) would likely be challenging in unscripted conversation because the measure of interest—lexical competition—would likely be eliminated by conversational constraints. Progress can be made through the pairing of traditional, well-controlled studies, with those conducted in more naturalistic settings. Insights and basic observations can be made on the basis of studies of natural conversation, which can then be tested in more controlled settings using standard paradigms. In cases where findings from natural conversation do not extend to scripted settings, further experimentation can identify the necessary conditions to observe the phenomena of interest, thereby informing the mechanisms involved. Blended methods, such as situations in which conversations are partially scripted (e.g., Brown-Schmidt, 2012), or in which the participant speaks with a confederate participant (e.g., Hanna & Tanenhaus, 2004; cf. Kuhlen & Brennan, 2013) are likely to be particularly useful in such situations as they afford control of key features of the interaction, while allowing other features of the interaction to unfold naturally.

In short, language use is fundamentally altered by conversational context, and as I have argued, conversation is the most basic site of language use. As a result,

building a general theory of language processing will *require* extensive study of language processing in unscripted conversational settings. While studies of language use in conversation are irreplaceable, significant advances in our understanding of language processing will also continue to require carefully controlled experiments in non-interactive settings. Whether the results of these experiments extend to language processing in conversation can subsequently be investigated using conversational paradigms. Finally, in addition to being a test-bed for the generalizability of results from standard paradigms, conversational studies can also be used as a tool for observing and documenting novel phenomena, which can then be studied more carefully in controlled settings, creating a feedback loop between the two approaches.

Acknowledgments

Preparation of this chapter was supported by National Science Foundation Grants NSF BCS 10-19161 and NSF 12-57029 to Sarah Brown-Schmidt.

References

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition: evidence for continuous mapping models. *Journal of Memory and Language, 38*, 419-439.

Altmann, G.T.M., & Kamide, Y. (2009). Discourse-mediation of the mapping between language and the visual world: Eye movements and mental representation. *Cognition, 111*, 55-71.

Arnold, J. E., Tanenhaus, M. K., Altmann, R. J., & Fagnano, M. (2004). The old and Thee, uh, new: Disfluency and reference resolution. *Psychological Science, 15*, 578-582.

Bailenson, J. N., Yee, N. (2005). Digital chameleons—automatic assimilation of nonverbal gestures in immersive virtual environments. *Psychological Science, 16*, 814–819.

Baldwin, D.A. (1991). Infants' contribution to the achievement of joint reference. *Child Development*, *62*, 875-890.

Baldwin, D.A. (1993). Early referential understanding: Infants' ability to recognize referential acts for what they are. *Developmental Psychology*, *29*, 832-843.

Bangerter, A. (2004). Using pointing and describing to achieve joint focus of attention in dialog. *Psychological Science, 15*, 415-419.

Barr, D. J. (2008). Pragmatic expectations and linguistic evidence: Listeners anticipate but do not integrate common ground. *Cognition, 109*, 18-40.

Becker-Asano, C. & Wachsmuth, I. (2010). Affective computing with primary and secondary emotions in a virtual human. *Journal of Autonomous Agents and Multi-Agent Systems, 20*, 32-49.

Beun, R.-J., & Cremers, A. H. M. (1998). Object reference in a shared domain of conversation. *Pragmatics & Cognition, 6*, 121-151.

Branigan, H. P., Pickering, M. J., & Cleland, A. A. (2000). Syntactic co-ordination in dialog. *Cognition, 75*, B13-25.

Branigan, H.P., Catchpole, C., & Pickering, M.J. (2011). What makes dialogues easy to understand? *Language and Cognitive Processes 26, 1667-1686*.

Brennan, S. E. (2005). How conversation is shaped by visual and spoken evidence. In J. Trueswell & M. Tanenhaus (Eds.), *Approaches to studying world-situated language use: Bridging the language-as-product and language-action traditions* (pp. 95-129). Cambridge, MA: MIT Press.

Brennan, S. E., Chen, X., Dickinson, C., Neider, M., & Zelinsky, G. (2008). Coordinating cognition: The costs and benefits of shared gaze during collaborative search. *Cognition, 106*, 1465-1477.

Brown-Schmidt, S. (2005). Language Processing in Conversation. Doctoral dissertation, University of Rochester.

Brown-Schmidt, S., Gunlogson, C., & Tanenhaus, M. K. (2008). Addressees distinguish shared from private information when interpreting questions during interactive conversation. *Cognition, 107*, 1122-1134.

Brown-Schmidt, S., & Hanna, J. E. (2011). Talking in another person's shoes: Incremental perspective-taking in language processing. *Dialog and Discourse, 2, 11-33.*

Brown-Schmidt, S. & Tanenhaus, M. K. (2008). Real-time investigation of referential domains in unscripted conversation: a targeted language game approach. *Cognitive Science, 32*, 643–684.

Brown-Schmidt, S. (2012). Beyond common and privileged: Gradient representations of common ground in real-time language use. *Language and Cognitive Processes, 27, 62-89.*

Brown-Schmidt, S. (2009a). Partner-specific interpretation of maintained referential precedents during interactive dialog. *Journal of Memory and Language, 61*, 171-190.

Brown-Schmidt, S. (2009b). The role of executive function in perspective-taking during on-line language comprehension. *Psychonomic Bulletin and Review, 16*, 893-900.

Brown-Schmidt, S. & Fraundorf, S. (submitted). Interpretation of informational questions modulated by joint knowledge and intonational contours.

Butterworth, G. & Itakura, S. (2000). How the eyes, head and hand serve definite reference. *British Journal of Developmental Psychology*, *18*, 25-50.

Caron, A.J., Butler, S., & Brooks, R. (2002). Gaze following at 12 and 14 months: Do the eyes matter? *British Journal of Developmental Psychology*, *20*, 225-239.

Chambers, C. G., & San Juan, V. (2008). Perception and presupposition in real-time language comprehension: Insights from anticipatory processing. *Cognition, 108*; 26-50.

Chambers, C. G., Tanenhaus, M. K., Eberhard, K. M., Filip, H., & Carlson, G. N. (2002). Circumscribing referential domains during real-time language comprehension. *Journal of Memory and Language, 47*, 30–49.

Chartrand, T. L., & Bargh, J. A. (1999). The Chameleon Effect: The perception-behavior link and social interaction. *Journal of Personality and Social Psychology, 76*, 893-910.

Clark, H. H. (1992). *Arenas of Language Use*. Chicago: University of Chicago Press.

Clark, H. H. (1996). *Using Language*. Cambridge: Cambridge University Press.

Clark, H. H., and Brennan, S. A. (1991). Grounding in communication. In L.B. Resnick, J.M. Levine, & S.D. Teasley (Eds.). *Perspectives on socially shared cognition*. Washington: APA Books.

Clark, H. H., & Krych, M. A. (2004). Speaking while monitoring addressees for understanding. *Journal of Memory and Language, 50*, 62-81.

Clark, H. H., & Marshall, C. R. (1978). Reference diaries. *Theoretical issues in natural language processing* (Vol. 2), ed. D. L. Waltz, 57-63. New York: Association for Computing Machinery.

Clark, H.H., & Marshall, C.R. (1981). Definite reference and mutual knowledge. *Elements of Discourse Understanding*, eds. A. K. Joshi, B. L. Webber, I. A. Sag, 10-63, Cambridge University Press.

Clark, H. H. & Schaefer, E. F. (1989). Contributing to discourse. *Cognitive Science, 13*, 259-294.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition, 22*, 1-39.

Cohen, P. (2010). Next Big Thing in English: Knowing They Know That You Know. March 31, 2010, *The New York Times*.

Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology, 6,* 84-107.

Creel, S. C., Aslin, R. N., & Tanenhaus, M.K. (2008). Heeding the voice of experience: The role of talker variation in lexical access. *Cognition, 106,* 63-664.

Creel, S. C. (2014). Preschoolers' flexible use of talker information during word learning. *Journal of Memory and Language, 73,* 81-98.

Dahan, D., Magnuson, J.S., Tanenhaus, M.K., & Hogan, E.M. (2001). Subcategorical mismatches and the time course of lexical access: Evidence for lexical competition. *Language and Cognitive Processes, 16*, 507–534.

Dahan, D., & Tanenhaus, M. K. (2004). Continuous mapping from sound to meaning in spoken-language comprehension: Immediate effects of verb-based

thematic constraints. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 498-513.

Dahan, D., Drucker, S. J., & Scarborough, R. A. (2008). Talker adaptation in speech perception: Adjusting the signal or the representations? *Cognition*, *108*, 710-718.

Deák, G.O., Flom, R.A. & Pick, A.D. (2000). Effects of gesture and target on 12- and 18-Month-Olds' joint visual attention to objects in front of or behind them. *Developmental Psychology*, *36*, 511-523.

DeLoache, J. S., Chiong, C., Sherman, K., Islam, N., Vanderborght, M., Troseth, G. L., Strouse, G. A., & O'Doherty, K. (2010). Do babies learn from baby media? *Psychological Science, 21*, 1570-1574.

Eberhard, K. M., Spivey-Knowlton, M. J., Sedivy, J. C., & Tanenhaus, M. K. (1995). Eye-movements as a window into spoken language comprehension in natural contexts. *Journal of Psycholinguistic Research, 24*, 409-436.

Ferguson, H. J. & Breheny, R. (2012). Listeners' eyes reveal spontaneous sensitivity to others' perspectives. *Journal of Experimental Social Psychology, 48,* 257-263.

Foxtree, J. E. (1999). Listening in on monologues and dialogues. *Discourse Processes, 27*, 35-53.

Garnsey, S. M., Pearlmutter, N. J., Meyers, E., & Lotocky, M. A. (1997). The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *Journal of Memory and Language, 37,* 58-93.

Garrod, S., & Anderson, A. (1987). Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition, 27*, 181-218.

Gergle, D., Kraut, R. E., & Fussell, S. R. (2004a). Action as language in a shared visual space. In *Proceedings of Computer Supported Cooperative Work (CSCW 2004)*, pp. 487-496. New York: ACM Press.

Gergle, D., Kraut, R. E., & Fussell, S. R. (2004b). Language efficiency and visual technology: Minimizing collaborative Effort with Visual Information. *Journal of Language and Social Psychology*, *23*, 491-517.

Gleitman, L. R., & Gleitman, H. (1992). A picture is worth a thousand words, but that's the problem: The role of syntax in vocabulary acquisition. *Current Directions in Psychological Science, 1*, 31–35.

Greene, S. B., Gerrig, R.J., McKoon, G., & Ratcliff, R. (1994). Unheralded pronouns and management by common ground. *Journal of Memory and Language, 33*, 511-526.

Griffin, Z. M., & Bock, K. (2000). What the eyes say about speaking. *Psychological Science, 11*, 274-279.

Grosz, B., & Sidner, C. (1986). Attentions, intentions and the structure of discourse. *Computational Linguistics, 12*, 175-204.

Hanna, J. E., & Brennan, S. E. (2007). Speakers' eye gaze disambiguates referring expressions early during face-to-face conversation. *Journal of Memory and Language, 57,* 596–615.

Hanna, J. E., & Tanenhaus, M. K. (2004). Pragmatic effects on reference resolution in a collaborative task: evidence from eye movements. *Cognitive Science, 28*, 105-115.

Hanna, J. E., Tanenhaus, M. K., & Trueswell, J. C. (2003). The effects of common ground and perspective on domains of referential interpretation. *Journal of Memory and Language, 49,* 43-61.

Haywood, S. L., Pickering, M. J., & Branigan, H. P. (2005). Do speakers avoid ambiguities during dialog? *Psychological Science, 16*, 362-366.

Heller, D., Grodner, D., & Tanenhaus, M. K. (2008). The role of perspective in identifying domains of reference. *Cognition, 108,* 831-836.

Heller, D., Parisien, C., & Stevenson, S. (2012). *Perspective-taking behav- ior as the probabilistic weighing of multiple domains*. Poster presented at the City University of New York Conference on Human Sentence Processing, New York, NY.

Horton, W. S. (2007). The influence of partner-specific memory associations on language production: Evidence from picture naming. *Language and Cognitive Processes, 22*, 1114-1139.

Horton, W.S., & Gerrig, R.J. (2005a). Conversational Common Ground and Memory Processes in Language Production. *Discourse Processes, 40*, 1-35.

Horton, W.S., & Gerrig, R.J. (2005b). The impact of memory demands on audience design during language production. *Cognition, 96*, 127-142.

Irwin, D. E. (2004). Fixation location and fixation duration as indices of cognitive processing. In J. Henderson & F. Ferreira (Eds.) *The interface of language, vision and action: Eye movements and the visual world*, pp. 105-133.

Kamide, Y., Altmann, G. T. M., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language, 49*, 133-156.

Kendon, A. (1970). Movement coordination in social interactions. *Acta Psychologica, 32*, 101–125.

Keysar, B., Lin, S., & Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition, 89*, 25-41.

Konopka, A. E., & Brown-Schmidt, S. (2014). Message Encoding. In: V. Ferreira, M. Goldrick, and M. Miozzo (Eds.), The Oxford Handbook of Language Production, Oxford University Press, New York, NY, pp. 1-20.

Kuhl, P. K., Tsao, F.-M., & Liu, H.-M. (2003). Foreign-language experience in infancy: Effects of short-term exposure and social interaction on phonetic learning. *Proceedings of the National Academy of Sciences, 100*, 9096-9101.

Kuhlen, A. K., & Brennan, S. E. (2010). Anticipating distracted addressees: How speakers' expectations and addressees' feedback influence storytelling. *Discourse Processes, 47*, 567-587.

Kuhlen, A. K. & Brennan, S. E. (2013). Language in dialogue: when confederates might be hazardous to your data. *Psychonomic Bulletin & Review, 20,* 54-72.

LaFrance, M. (1979). Nonverbal synchrony and rapport: Analysis by the cross-lag panel technique. *Social Psychology Quarterly, 42*, 66–70.

LaFrance, M., & Broadbent, M. (1976). Group rapport: Posture sharing as a nonverbal indicator. *Group and Organizational Studies, 1*, 328–333.

Landragin, F. (2006). Visual perception, language and gesture: A model for their understanding in multimodal dialog systems. *Signal Processing, 86*, 3578-3595.

Levelt, W. J. M., & Kelter, S. (1982). Surface form and memory in question answering. *Cognitive Psychology, 14*, 78-106.

Levy, E. T., & McNeill, D. (1992). Speech, gesture, and discourse. *Discourse Processes 15*, 277-301.

Lockridge, C.B., & Brennan, S. E. (2002). Addressees' needs influence speakers' early syntactic choices. *Psychonomic Bulletin & Review,* 9, 550-557.

McMurray, B., Aslin, R. N., Tanenhaus, M. K., Spivey, M. J., & Subik, D. (2008). Gradient sensitivity to within-category variation in words and syllables. *Journal of Experimental Psychology: Human Perception and Performance, 34*, 1609-1631.

McRae, K., Hare, M., & Tanenhaus, M.K. (2005). Meaning Through Syntax is insufficient to explain comprehension of sentences with reduced relative clauses: A critique of McKoon & Ratcliff (2003). Psychological Review, 112, 1022-1031.

Morales, M., Mundy, P., & Rojas, J. (1998). Following the direction of gaze and language development in 6-month olds. *Infant behavior and development*, *21*, 373-377.

Morales, M., Mundy, P., Delgado, C.E.F, Yale, M., Neal, R., & Schwartz, H.K. (2000). Gaze following, temperament, and language development in 6-month-olds: A replication and extension. *Infant behavior and development*, *23*, 231-236.

Moses, L.J., Baldwin, D.A., Rosicky, J.G., & Tidball, G. (2001). Evidence for referential understanding in the emotions domain at twelve and eighteen months. *Child Development*, *72*, 718-735.

Nadig, A.S. & Sedivy, J.C. (2002). Evidence of perspective-taking constraints in children's on-line reference resolution. *Psychological Science, 13*, 329-336.

National Center for Education Statistics (2003). http://nces.ed.gov/naal/kf_demographics.asp

Neider, M. B., Chen, X., Dickinson, C. A., Brennan, S. E., & Zelinsky, G. J. (2010). Coordinating spatial referencing using shared gaze. *Psychonomic Bulletin and Review, 17*, 718-724.

Nurmsoo, E. & Bloom, P. (2008). Preschoolers' perspective taking in word learning: Do they blindly follow eye gaze? *Psychological Science, 19*, 211-215.

Olson, D. R. (1970). Language and thought: Aspects of a cognitive theory of semantics. *Psychological Review, 77*, 257–273.

Osgood, C. E. (1971). Where do sentences come from? In D. D. Steinberg & L. A. Jakobovits (Eds.), *Semantics: An interdisciplinary reader in philosophy, linguistics and psychology* (pp. 497–529). Cambridge, MA: Cambridge University Press.

Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *Journal of the Acoustical Society of America, 119*, 2382-2393.

Pechmann, T. (1989). Incremental speech production and referential overspecification. *Linguistics, 27*, 89–110.

Pfeiffer-Leßmann, N., & Wachsmuth, I. (2009). Formalizing joint attention in cooperative interaction with a virtual human. In B. Mertsching, M. Hund, & Z. Aziz (Eds.), *KI 2009: Advances in Artificial Intelligence* (pp. 540-547). Berlin: Springer (LNAI 5803).

Pickering, M. J., Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences 27*, 169–225.

Poesio, M. & Rieser, H. (2010). Completions, coordination, and alignment in dialog. *Dialog and Discourse, 1,* 1-89.

Purver, M. & Kempson, R. (2004). Incrementality, Alignment and Shared Utterances. In: *Catalog '04: Proceedings of the Eighth Workshop on the Semantics and Pragmatics of Dialogue,* J. Ginzburg & E. Vallduví, Eds., Barcelona, pp. 85-92.

Reitter, D. & Moore, J. D. (2007). Predicting success in dialogue. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL)*, pages 808-815, Prague, Czech Republic.

Reitter, D., Moore, J. D., & Keller, F. (2006). Priming of syntactic rules in task-oriented dialogue and spontaneous conversation. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society (CogSci)*, pages 685-690, Vancouver, Canada, 2006.

Richardson, D. C., & Dale, R. (2005). Looking to understand: The coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension. *Cognitive Science, 29*, 1045–1060.

Richardson, D. C., Dale, R., & Kirkham, N. Z. (2007). The art of conversation is coordination: Common ground and the coupling of eye movements during dialogue. *Psychological* Science, 18, 407-413.

Richardson, D. C., Dale, R., & Tomlinson, J. M. (2009). Conversation, Gaze Coordination, and Beliefs About Visual Context. *Cognitive Science, 33*, 1468-1482.

Roque, A. & Traum, D. (2008, June). Degrees of Grounding Based on Evidence of Understanding. In: *Proceedings of The 9th SIGdial Workshop on Discourse and Dialogue (SIGdial 2008)*, Columbus, OH.

Roque, A. & Traum, D. (2009, July). Improving a Virtual Human Using a Model of Degrees of Grounding. In: *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI-09)*, Pasadena, CA.

Rubin, R. D., Brown-Schmidt, S., Duff, M. C., Tranel, D., & Cohen, N. J. (2011). How do I remember that I know you know that I know? *Psychological Science, 22*, 1574-1582.

Ryskin, R.A., Brown-Schmidt, S., Canseco-Gonzalez, E., Yiu, E.K., & Nguyen, E.T. (2014). Visuospatial Perspective-taking in Conversation and the Role of Bilingual Experience. *Journal of Memory and Language, 74,* 46-76*.*

Scaife, M. & Bruner, J.S. (1975). The capacity for joint visual attention in the infant. *Nature*, *253*, 265-266.

Schegloff, E. A. (1984). On Some Gestures' Relation to Talk. In J. M. Atkinson and J. Heritage (eds.), *Structures of Social Action*, (Cambridge: Cambridge University Press, 1984), 266-298.

Schober, M. F. (1993). Spatial perspective-taking in conversation. *Cognition, 47,* 1-24.

Schober, M. F., & Clark, H. H. (1989). Understanding by addressees and overhearers. *Cognitive Psychology, 21*, 211-232.

Simons, D. J., & Chabris, C. F. (1999). Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception, 28*, 1059-1074.

Speer, N. K., & Zacks, J. M. (2005). Temporal changes as event boundaries: Processing and memory consequences of narrative time shifts. *Journal of Memory and Language, 53,* 125-140.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science, 268*, 1632-1634.

Tesink, C.M.J.Y., Petersson, K., M., van Berkum, J. J. A., van den Brink, D., Buitelaar, J. K., & Hagoort, P. (2008). Unification of speaker and meaning in language comprehension: An fMRI study. *Journal of Cognitive Neuroscience, 21:11*, 2085-2099.

Thothathiri, M. & Snedeker, J. (2008). Give and take: Syntactic priming during spoken language comprehension. *Cognition, 108,* 51-68.

Trude, A. M., & Brown-Schmidt, S. (2012). Talker-specific perceptual adaptation during on-line speech perception. *Language and Cognitive Processes, 27, 979-1001.*

Trueswell, J. C., & Tanenhaus, M. K. (1994). Toward a lexicalist framework for constraint-based syntactic ambiguity resolution. *Perspectives on Sentence Processing*, eds. C. Clifton, L. Frazier, & K. Rayner, 155-179. Lawrence Erlbaum Assoc.

US Dept of Labor, (2010). American Time Use Survey, 2009. http://www.bls.gov/news.release/atus.htm

Van Berkum, J. J. A., Van den Brink, D., Tesink, C. M. J. Y., Kos, M., & Hagoort, P. (2008). The neural integration of speaker and message. *Journal of Cognitive Neuroscience, 20*, 580-591.

Wachsmuth, I. (2008). 'I, Max' – Communicating with an Artificial Agent. In: I. Wachsmuth and G. Knoblich (Eds.): Modeling Communication, LNAI 4930, pp. 279–295, 2008. Springer-Verlag Berlin Heidelberg.

Wilkes-Gibbs, D., & Clark, H. H. (1992). Coordinating beliefs in conversation. *Journal of Memory and Language, 31*, 183-194.

Wilson, M. P., & Garnsey, S. M. (2009). Making simple sentences hard: Verb bias effects in simple direct object sentences. *Journal of Memory and Language, 60*, 368-392.

Yoon, S.O. & Brown-Schmidt, S. (2014). Adjusting conceptual pacts in three-party conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*, 919-937.

Yu, C., Ballard, D.H. & Aslin, R.N. (2005). The Role of Embodied Intention in Early Lexical Acquisition, *Cognitive Science, 29*, 961-1005.

Zacks, J. M. (2004). Using movement and intentions to understand simple events. *Cognitive Science, 28,* 979-1008.